When are Adaptive Designs really needed?

Christopher Jennison

Dept of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

European Clinical Research Infrastructure Network

International Clinical Trials' Day

Luxembourg, May 2014

Outline of talk

1. The traditional approach to Phase III clinical trial design

   Keeping Phase III trials simple

   Early adaptive methods and group sequential designs

2. Recent developments in adaptive Phase III designs

   Combining data across stages

   Testing multiple hypotheses

3. Case study

   A clinical trial with a survival endpoint and treatment selection

4. Conclusions

1. The role of a Phase III clinical trial

Phase III trials are conducted at the end of the drug development process, or the development of a new medical treatment. Then,

> The treatment has been refined and tested in earlier development and in Phase I and II trials,
>
> A substantial body of work supports the investigators’ belief that the new treatment is effective and safe.

The aim of the Phase III trial is to compare the new treatment with the current standard treatment or a placebo, when given to the target patient population.

The need for a clear, unambiguous comparison leads to the desire for a simple Phase III clinical trial.

All aspects of the Phase III trial design are pre-defined and written into the protocol and statistical analysis plan.

Traditional Phase III clinical trials (pre 2000 approx.)

A trial protocol specifies:

The experimental treatment and the control or placebo treatment,

The patient population (eligibility criteria, etc.),

Sample size for the trial,

Statistical analysis plan.

Interim analyses may be conducted to:

Monitor safety,

Stop early for futility if the new treatment is not effective,

Stop early if there is overwhelming evidence of efficacy.

Many of the decisions taken in creating such a design would benefit from further knowledge of the treatment, the patients, or patient responses.

Early examples of adaptive methods

There is a long history of “Adaptive” statistical methods.

Adaptive randomisation

In a trial comparing two treatments, adaptive randomisation can be used to increase the proportion of patients allocated to the better of two treatments.

However, once randomisation becomes unequal, ethical issues may arise as to whether it is permissible to randomise at all.

Adaptive randomisation highlights the role of “equipoise” in a randomised clinical trial and just what this term should mean.

Ethical and statistical concerns were clearly evident in two Harvard trials in the 1970s and 1980s which investigated ECMO treatment of critically ill, new-born babies (Ware, *Statistical Science*, 1989).

Early examples of adaptive methods

Sample size re-estimation

The sample size needed to achieve a specific power under a given treatment effect is proportional to the response variance — which is typically unknown when planning a trial.

Wittes & Brittain (*Statistics in Medicine*, 1990) suggested choosing an initial sample size based on a plausible response variance, then updating the sample size as better estimates of response variance are obtained.

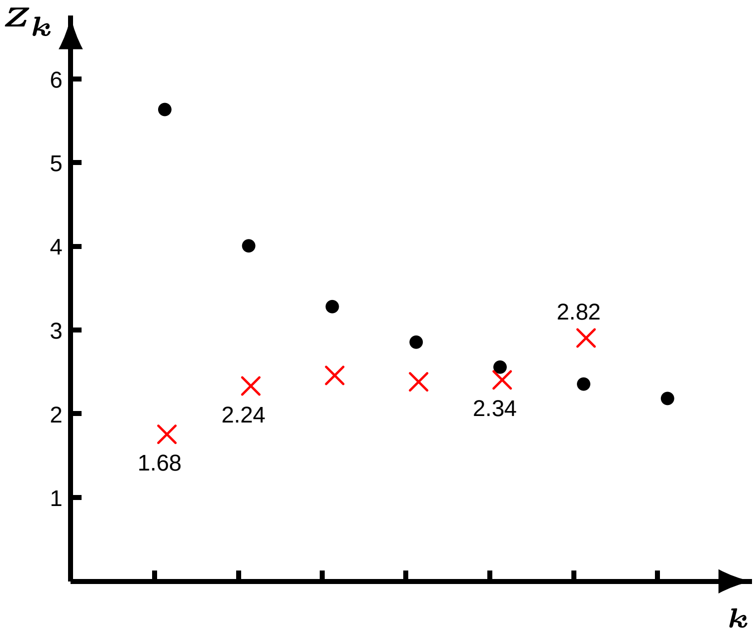
The same approach can be used to handle an unknown baseline hazard rate for survival data.

Sample size re-estimation in the light of estimates of “nuisance parameters” is still one of the most commonly used adaptive methodologies.

Sequential analysis of clinical trials

Group sequential methods, introduced in the late 1970s, allow early stopping for either a positive or negative final decision.

An early example, the the Beta-Blocker Heart Attack Trial, compared propanolol with placebo. (DeMets et al., *Controlled Clinical Trials*, 1984)



The trial stopped with a positive outcome after the 6th of 7 planned analyses.

Sequential analysis of clinical trials

In their book *Group Sequential Methods with Applications to Clinical Trials*, Jennison & Turnbull (2000) gave a unified treatment of group sequential methods, including:

General theory of group sequential analysis

Early stopping for futility or for a positive outcome

Survival data

Error spending designs that adapt to unpredictable information levels

Sample size re-estimation as nuisance parameters (but not the treatment effect) are estimated

Multiple endpoints or multiple treatments

Although some key papers started to appear in the mid 1990s, Jennison & Turnbull did not discuss modern adaptive methods.

2. Adaptive designs for Phase III clinical trials

We noted that many of the decisions taken in designing a clinical trial would benefit from further knowledge of the treatment, the patients, or patient responses.

What is the best dose for the new treatment?

What is the best method of delivery for the new treatment?

Does the treatment have greater benefit for a sub-population of patients?

For a normally distributed response, what is the variance?

Or, for time-to-event data, what is the baseline hazard rate?

How large a treatment effect is clinically significant?

How large a treatment effect is anticipated?

Such questions are addressed throughout the development of a new treatment.

Adaptive designs allow final changes to be made as new information is gathered during a Phase III trial.

Adaptive designs for Phase III clinical trials

In the early 2000s, industry and regulators were aware of falling success rates in late stage trials — a “statistical” solution would be very welcome indeed!

The idea of shifting from rigidly defined Phase III clinical trials to a flexible, adaptive approach was both attractive and challenging.

Sceptics asked:

> Can the results of an adaptive trial be statistically valid and credible?
>
> Will regulators accept adaptive designs?
>
> What features of a trial should be adapted?
>
> What are the benefits of adaptation?

Some proposals seemed to violate fundamental statistical principles.

There was a need for critical appraisal of new methodologies.

3. The statistical building blocks of adaptive clinical trials

(i) Testing a null hypothesis by combining data across stages

A key piece of methodology for hypothesis testing in adaptive designs is the ***Combination test*** (Bauer & Köhne, *Biometrics*, 1994).

Initial design

Define the null hypothesis, H_0: $\theta \leq 0$, and say a combination test will be used.

Design Stage 1, fixing sample size and test statistic for this stage.

Stage 1

Observe P_1, the one-sided P-value for testing H_0 based on Stage 1 data.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe P_2, the one-sided P-value for testing H_0 based on Stage 2 data.

Under $\theta = 0$: $P_1 \sim U(0,1)$, $P_2 \sim U(0,1)$, and P_1 and P_2 are independent.

Bauer & Köhne's inverse $\boldsymbol{\chi^2}$ combination test

The inverse $\boldsymbol{\chi^2}$ test rejects H_0 for low values of $P_1\, P_2$.

If $P \sim U(0,1)$, then $-\ln(P) \sim$ Exp $(1) = \frac{1}{2}\,\chi^2_2$.

Thus, under $\theta = 0$,

$$-\ln(P_1\, P_2) \;\sim\; \frac{1}{2}\,\chi^2_4.$$

Combining the two P-values in an overall test, we reject H_0 if

$$-\ln(P_1\, P_2) \;>\; \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

Despite the data-dependent adaptation, the overall type I error rate is still protected at level $\boldsymbol{\alpha}$ under $\boldsymbol{H_0}$.

This χ^2 test was originally proposed for combining results of several studies by R. A. Fisher (1932) *Statistical Methods for Research Workers*.

Methods for combining data across stages of an adaptive trial

Other forms of combination test (Bauer & Köhne, 1994) are available, such as the “inverse normal” combination rule.

Or, methods can be based on preserving the conditional type I error probability:

Proschan & Hunsberger (*Biometrics*, 1995)

Denne (*Statistics in Medicine*, 2001)

Müller & Schäfer (*Biometrics*, 2001 and *Statistics in Medicine*, 2004)

L. D. Fisher (*Statistics in Medicine*, 1998) proposed a “variance spending” approach.

Adaptation can occur in a group sequential clinical trial:

Cui, Hung & Wang (*Biometrics*, 1999)

Lehmacher & Wassmer (*Biometrics*, 1999)

Despite their varied descriptions and derivations, there is much in common between all of these methods.

Adaptive designs using a combination test

Let θ denote the treatment effect, e.g., the difference in mean response between patients on the new treatment and patients on control.

We test H_0: $\theta \leq 0$ vs $\theta > 0$, where $\theta > 0$ means the new treatment is superior.

A combination test may be used when sample size is re-estimated in response to a new estimate for a nuisance parameter, or an estimate of θ itself.

There is no problem of bias in the overall type I error rate caused by this sample size re-estimation.

Suppose the experimental treatment is modified in mid-study

We could use a combination test of H_0: "$\theta \leq 0$ *and* $\tilde{\theta} \leq 0$", where $\tilde{\theta}$ denotes the treatment effect of the modified treatment.

On rejecting H_0 we conclude either $\theta > 0$ *or* $\tilde{\theta} > 0$ — but we cannot say which.

With multiple hypotheses, further tools are needed . . .

The statistical building blocks of adaptive clinical trials

(ii) Testing multiple hypotheses

There may be reasons to change the null hypothesis or to select from a set of possible null hypotheses during a clinical trial:

Selecting one out of several versions of a treatment,

Restriction to a sub-group of the patient population,

Switching from a test of superiority to a test of non-inferiority.

When H_0 changes or is selected, attention focuses on the ***new*** null hypothesis. Care is needed to avoid ***selection bias*** as this hypothesis is data-generated.

Methods must protect the type I error rate when there are multiple hypotheses.

Selected references:

Bauer & Köhne (*Biometrics*, 1994), Bretz, Schmidli et al. (*Biometrical Journal*, 2006), Schmidli, Bretz et al. (*Biometrical Journal*, 2006).

Testing multiple hypotheses

The familywise error rate

Suppose there are h null hypotheses, H_i: $\theta_i \leq 0$ for $i = 1, \ldots, h$.

A procedure's ***familywise error rate*** under a set of values $(\theta_1, \ldots, \theta_h)$ is

$$P\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\} \;=\; P\{\text{Reject any true } H_i\}.$$

The familywise error rate is controlled ***strongly*** at level α if this error rate is at most α for all possible combinations of θ_i values. Then

$$P\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } (\theta_1, \ldots, \theta_h).$$

Using such a procedure, the probability of choosing to focus on a parameter θ_{i*} and then falsely claiming significance for null hypothesis H_{i*} is at most α.

Testing multiple hypotheses: Closed testing procedures

Marcus et al. (*Biometrika*, 1976) define a ***closed testing procedure*** which combines level α tests of each H_i and of intersections of these hypotheses.

We have null hypotheses H_i, $i = 1, \ldots, h$.

For each subset I of $\{1, \ldots, h\}$, define the intersection hypothesis

$$H_I = \cap_{i \in I} H_i.$$

Construct a level α test of each intersection hypothesis H_I, i.e., a test which rejects H_I with probability at most α whenever all hypotheses specified in H_I are true.

Closed testing procedure

The simple hypothesis H_j: $\theta_j \leq 0$ is rejected overall if, and only if, H_I is rejected for every set I containing index j.

It can be show (quite easily) that this procedure provides strong control, at level α, of the familywise error rate.

Putting the building blocks together

Closed testing procedures can be used to test multiple hypotheses in a single stage, non-adaptive design.

One may wish to test hypotheses about secondary endpoints or patient sub-groups after obtaining a positive result on the primary endpoint.

Positive results will be included in the labelling of the new treatment.

When several null hypotheses arise in a group sequential or adaptive trial

In constructing a closed testing procedure, we need to define a combination test for each simple hypothesis and each intersection of simple hypotheses.

Each of these tests will combine data across stages of the trial.

The key requirement is that, for each hypothesis test, we decide how the P value will be computed from data in the next stage *before that stage of the trial is carried out*.

3. Case study: A clinical trial with a survival endpoint and treatment selection

Consider a trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing

Experimental Treatment 2: Slower dosing

Control treatment

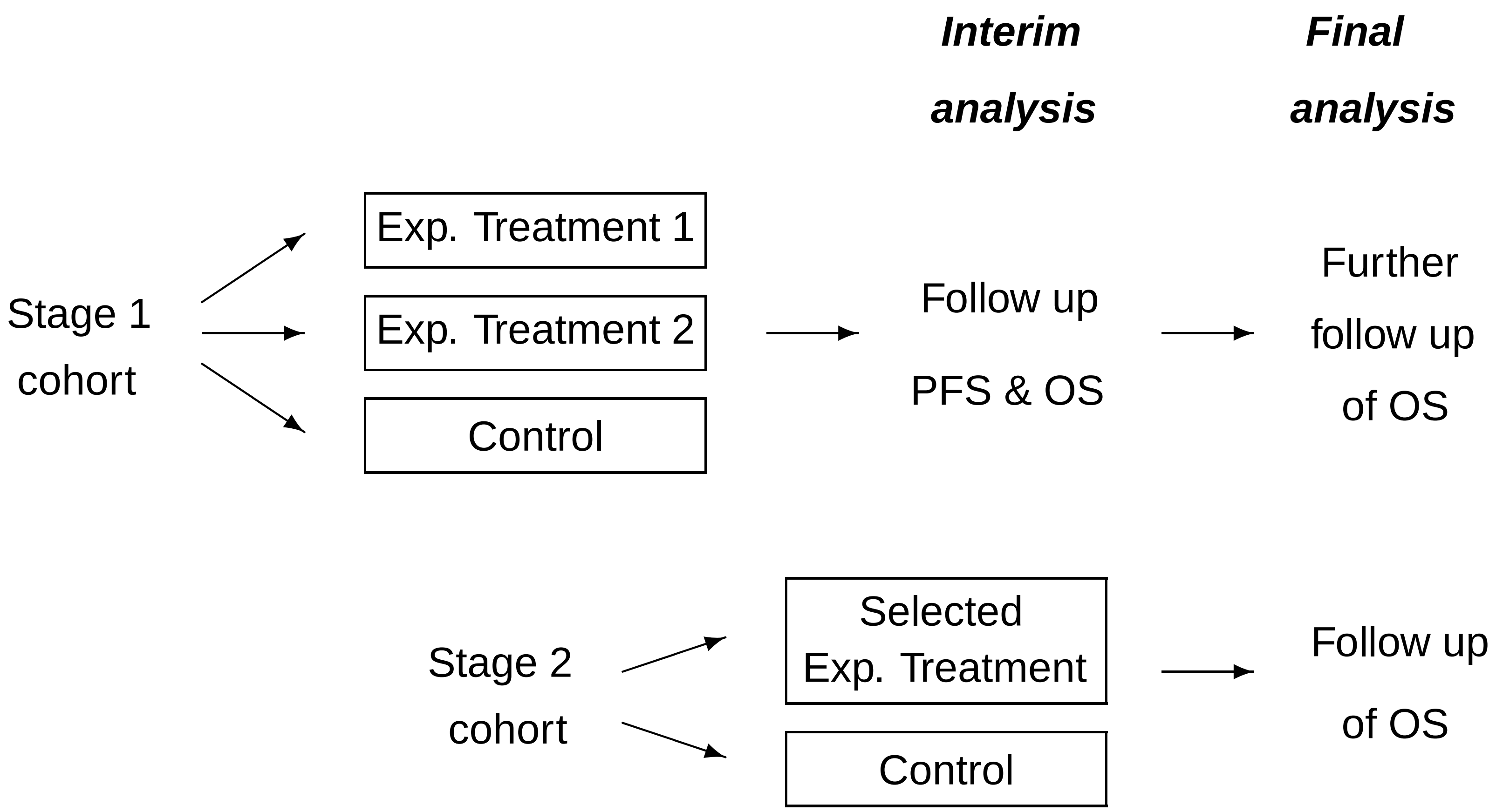
The primary endpoint is Overall Survival (OS).

Information on OS, Progression Free Survival (PFS) and safety will be used at an interim analysis to choose between the two experimental treatments.

Note that PFS is useful here as it is more rapidly observed.

After the interim analysis, patients will only be recruited to the selected treatment and the control.

Overall plan of the trial



Interim analysis

Final analysis

Stage 1 cohort

Exp. Treatment 1

Exp. Treatment 2

Control

Follow up

PFS & OS

Further follow up of OS

Stage 2 cohort

Selected Exp. Treatment

Control

Follow up of OS

At the final analysis, we test the null hypothesis that OS on the selected treatment is no better than OS on the control.

Protecting the type I error rate

We may assume a proportional hazards model with

$$\lambda_1 = \text{Hazard ratio, Control } \textit{vs} \text{ Exp. Treatment 1}$$

$$\lambda_2 = \text{Hazard ratio, Control } \textit{vs} \text{ Exp. Treatment 2}$$

$$\theta_1 = \log(\lambda_1), \quad \theta_2 = \log(\lambda_2).$$

We test null hypotheses

$$H_{0,1}\text{: } \theta_1 \leq 0 \quad \text{vs} \quad \theta_1 > 0 \quad \textit{(Exp. Treatment 1 superior to control)},$$

$$H_{0,2}\text{: } \theta_2 \leq 0 \quad \text{vs} \quad \theta_2 > 0 \quad \textit{(Exp. Treatment 2 superior to control)}.$$

We require

$$Pr\{\text{Reject any true null hypothesis}\} \leq \alpha.$$

A closed testing procedure

Define level α tests of

$$H_{0,1}\colon\ \theta_1 \leq 0,$$

$$H_{0,2}\colon\ \theta_2 \leq 0$$

and of the intersection hypothesis

$$H_{0,12} \ = \ H_{0,1} \cap H_{0,2}\colon\ \ \theta_1 \leq 0 \ \text{ and } \ \theta_2 \leq 0.$$

Then:

*Reject $H_{0,1}$ **overall** if the above tests reject $H_{0,1}$ and $H_{0,12}$,*

*Reject $H_{0,2}$ **overall** if the above tests reject $H_{0,2}$ and $H_{0,12}$.*

The requirement to reject $H_{0,12}$ compensates for testing multiple hypotheses and the “selection bias” in choosing the treatment to focus on in Stage 2.

Combination tests

In the closed testing procedure, each null hypothesis is tested using a combination test to combine P-values from the two stages.

Overall survival data within each stage are analysed using a logrank test.

In testing the intersection hypothesis $H_{0,12}$: $\theta_1 \leq 0$ and $\theta_2 \leq 0$, the logrank test can compare survival in the pooled Exp. Treatment 1 and Exp. Treatment 2 patients vs the Control group.

There is an elegant theory for the behaviour of logrank statistics based on the increasing follow-up of a group of subjects (Tsiatis, *Biometrika*, 1981).

However, this theory may not be applicable in an adaptive trial (Bauer & Posch, *Statistics in Medicine*, 2004).

The problem can be solved by changing the definitions of “Stage 1” and “Stage 2” data (Jenkins, Stone & Jennison, *Pharmaceutical Statistics*, 2011).

Jenkins, Stone & Jennison (2011)

In constructing a combination test, it is natural to separate data into the parts accrued before and after the interim analysis:

	P_1	P_2
Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)
Stage 2 cohort		Overall survival (during Stage 2)

To avoid bias in a combination test, divide the data into parts from the two cohorts:

Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)	P_1
Stage 2 cohort		Overall survival (during Stage 2)	P_2

P-values for combination tests

All patients in the Stage 1 cohort are followed for overall survival up to a fixed time or a pre-specified number of failures.

The “Stage 1” statistics are based on the final OS data for the Stage 1 cohort

$P_{1,1}$ from log-rank test of Exp Tr 1 vs Control

$P_{1,2}$ from log-rank test of Exp Tr 2 vs Control

$P_{1,12}$ from log-rank test of combined Exp Tr 1 and Exp Tr 2 vs Control.

The “Stage 2” statistics are based on OS data for the Stage 2 cohort

If Exp. Treatment 1 is selected:

$P_{2,1}$ from log-rank test of Exp Tr 1 vs Control, $P_{2,12} = P_{2,1}$

If Exp. Treatment 2 is selected:

$P_{2,2}$ from log-rank test of Exp Tr 2 vs Control, $P_{2,12} = P_{2,2}$.

4. Conclusions

1. Experience over 50 or 60 years led to reliable and well understood methodology for conducting Phase III trials.

2. Innovation had taken place but its potential impact was most probably lessened by cautious investigators and conservative regulators.

3. At the turn of the millennium, the announcement of adaptive methods caused great excitement and offered hope of a step change in the success rate of clinical trials.

4. We now have

   Clearer appraisal of the benefits of adaptive design,

   Practical experience of conducting adaptive trials.

5. Opinions may have changed as to where adaptation can be most valuable, but the outlook is still positive.